DETECTING DRIVER INATTENTION BY ROUGH ICONIC CLASSIFICATION

G. Masala and E. Grosso

July 30, 2012

University of Sassari

Computer Vision Laboratory

Technical Report No. CVL -2012-002

UNIVERSITY
of
SASSARI

DETECTING DRIVER INATTENTION BY ROUGH ICONIC CLASSIFICATION

G. Masala and E. Grosso

University of Sassari
Computer Vision Laboratory
Porto Conte Ricerche - Loc. Tramariglio
07041 Alghero (SS)
`www.uniss.it`

July 30, 2012

Technical Report No. CVL -2012-002

Abstract

The paper proposes an original method , derived from basic face recognition and classification research, which is a good candidate for an effective automotive application. The proposed approach exploits a single b/w camera, positioned in front of the driver, and a very efficient classification strategy, based on neural network classifiers.

A peculiarity of the work is the adoption of iconic data reduction, avoiding specific and time-consuming feature-based approaches. Though at an initial development stage, the method proved to be fast and robust compared to state of the art techniques; experimental results show real-time response and mean weighted accuracy near to 92%. The method requires a simple training procedure which can be certainly improved for real applications; moreover it can be easily integrated with techniques for automatic face-recognition of the driver.

Keywords

Automotive safety, Computer vision, Visual attention, Pattern classification, Neural networks.

1 INTRODUCTION

The effect of fatigue on driving performance has been widely studied from physiologists and transportation experts. Fatigue has been proved to be a main cause of road accidents and pushed automotive corporations, since late 90s, toward the development of on-board intelligent safety systems, useful to evaluate in real time the driver's state of vigilance [1].

Available studies [2] identify inattention (a single term including distraction and fatigue effects) as the primary cause of crashes. At least 25% of road accidents in Europe is estimated to be directly related to a low attention state of the driver; one half of these events is attributed to distraction, resulting in more than 200.000 injured people per year.

This complex picture is attracting the interest of the scientific community. In particular, in comparison to other approaches, dated and rather intrusive, computer vision techniques have been considered suitable to detect changes in the facial features which characterize behaviors of inattentive people. Typically, persons with reduced alertness due to fatigue show longer blink duration, slow eyelid movement, small degree of eye opening , frequent nodding, yawning and drooping posture [3]. In case of distraction, common situations include wrong gaze direction and persistent rotation of the head.

The design of a fully automated safety system based on computer vision can benefit of a number of robust tools, coming from basic image analysis and related fields like biometrics; in fact, for a long time research on face recognition has been focusing on the detection and processing of specific facial features [4]. However, the application on a moving vehicle presents new challenges like changing backgrounds and variable lighting. Moreover, a useful system should guarantee real time performance and quick adaptability to a variable set of users and to natural movements performed during driving.

This paper proposes an original method , derived from basic face recognition and classification research, which is a credible candidate for an effective automotive application. Using a single frontal camera, the approach allows the detection of the driver and the simple classification of each frame in two states (attentive, inattentive) based on a pre-learned scheme. The method is simple and robust compared to state of the art techniques; moreover it is fast and requires a very simple training procedure.

Following sections are organized as follows: section 2 discusses previous work while section 3 details the proposed methodology. Experimental results are described and discussed in section 4. Finally section 5 draws some conclusions and analyses possible improvements.

2 PREVIOUS WORK

Several computer vision techniques have been proposed to detect driver inattention [5,6]. A common processing scheme, well discussed in [7,8] includes the following steps:

- face localization;
- localization of facial features (e.g. eyes or mouth);
- estimation of specific cues related to fatigue or distraction;
- fusion of cues in order to determine the global attention level.

Very often the localization step is accompanied by a tracking process; this strategy in normal conditions guarantees a drastic increase of performance.

Concerning face localization, very robust techniques have been developed in late 90s based on neural networks [9,10]. In 2004 Viola and Jones [11] proposed a new algorithm based on integral images and robust classification that achieves very good results and guarantees high performance. Both these approaches belong to the image-based subclass of the face detection techniques. More recently also feature-base approaches demonstrated a reasonable level of efficiency. In particular Particle Swarm Optimization [12] has been proposed for locating and tracking a limited number on facial landmarks.

Concerning facial features, researchers are mainly focusing on eyes and mouth. Work on feature detection is generally based on classification [4]; in particular Gabor and SVM techniques have been successfully proposed [7]. In order to work under low light conditions, researchers proposed the use of infrared illuminators, exploiting high reflection of the pupils [5]; as noted in [13], however, IR based approaches show malfunctions during daytime and require the installation of additional hardware.

Head rotation can be estimated by applying both 2D and 3D approaches [11]; generally, 3D approaches are more robust but require multiple cameras and quite expensive registration techniques.

Most of the literature work defines the PERCLOS as the main cue for the estimation of driver's fatigue. PERCLOS is a measure of the time percentage during which eyes remain closed 80% or more; in order to compute this cue, evyer image frame is usually classified into two classes (closed eyes or open eyes). k-NN techniques, SVMs and Bayes approaches have been successfully applied to this purpose [7].

Concerning the adoption and the fusion of different cues, certainly the head pose represents the most interesting issue [8]; recent works also introduce eye blinking detection [13], slouching frequency and postural adjustment. In [14] eye-mouth occlusion and 3D gaze are detected and used separately; fusion based on fuzzy rules is described by [9] showing some limited increase of performance.

The main contribution of this paper is the demonstration that the adoption of complex cues, and the following fusion of these cues, can be efficiently replaced by generalizing the concept of "inattentive driver". This iconic generalization, derived by processing and classifying off-line a significant number of real sequences, produces, as in the Viola Jones face detector, a pre-learned pattern that can be usefully exploited for on-line processing, achieving high levels of accuracy and real time performance.

3 THE ATTENTION MODEL

To the aim of this work, we define as "inattentive driver" a subject showing distraction, fatigue effects or both. The automatic system is therefore expected to detect situations like closed eyes, nodding and rotation of the head.

In order to guarantee fast processing of the video sequences, the attention model is based on the three step procedure depicted in figure 1:

1. Regions of interest are first extracted by applying the Viola Jones face detector;
2. a Sanger neural network is used to reduce the dimensionality of the feature space;
3. a feed forward neural network classifier decides about the attentive/inattentive state of the driver.

Viola Jones face detector [11] relies on the use of simple Haar-like features. It generates a large set of features and uses the AdaBoost algorithm to reduce the over-complete set. The detector is applied to gray-scale images, producing a set of scaled windows containing face-candidates.

The second and the third steps use feed-forward back-propagation neural networks [15]. The minimal unit of a neural network is an artificial neuron, constituted by a sum module of his input and a following filter (i.e. a transfer function as a sigmoid). The neurons are organized in layers; neurons belonging to a given layer can be connected only to neurons belonging to preceding and subsequent layers. In the training phase, weights corresponding to neural connections are set: for a supervised system, the network is trained by using samples of known classes whilst for unsupervised systems the training is based on the minimization of a generic function of the data and the network's output.

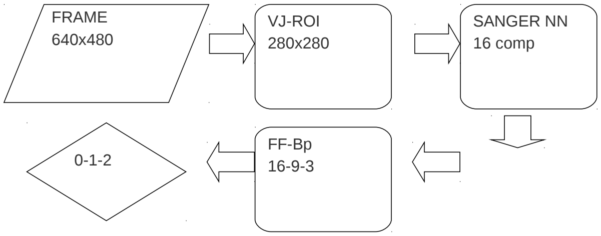

Figure 1: In the block diagram the input frame is first processed by the Viola-Jones algorithm; the resulting region of interest (ROI) is coded trough an Sanger neural network into a vector of 16 components. Finally, a feed forward neural network decides about the attention state of the driver.

A Sanger neural network [16] is a simple three-layer feed forward unsupervised neural network (with linear transfer function in the hidden neurons) which develops an internal representation corresponding to the principal components analysis of the full input data set. The network has three layers: input and output layers have the same dimension of the input patterns while the dimension of the hidden layer (corresponding to the number of the principal components) is determined during the training phase. The network is trained as an auto-encoder [15], in such a way to reproduce at the output the input data.

The feed forward neural network used for the third step is composed of 16 input neurons and 3 output neurons; note that the output neurons correspond to the three attention states considered in this work (attentive, inattentive-fatigued, inattentive-distracted). As discussed in the next section, the network is trained trough a back propagation algorithm on a training set and tested on a validation set to determine the optimal number of neurons in the hidden layer.

4 EXPERIMENTS

4.1 Data collection

In order to test the effectiveness of the proposed approach, a wi-fi pinhole camera has been installed in a car, as shown in figure 2. This camera allows the recording of several minutes of video during typical driving situations.

We collected data from 6 acquisition sessions of the same driver in different moments of the day and various conditions of ambient light. The user was driving both wearing glasses or not, without caring about the position of the seat and of the camera. Each session consists of 3 minutes of video recording, manually classified as follows :

- about one minute of normal driver behavior: the driver looks at the road straightaway or to rear view mirrors;

- about one minute of simulated fatigue effects: the driver closes the eyes and simulates nodding .

- about one minute of distracted behaviour; the driver looks up, down or laterally.

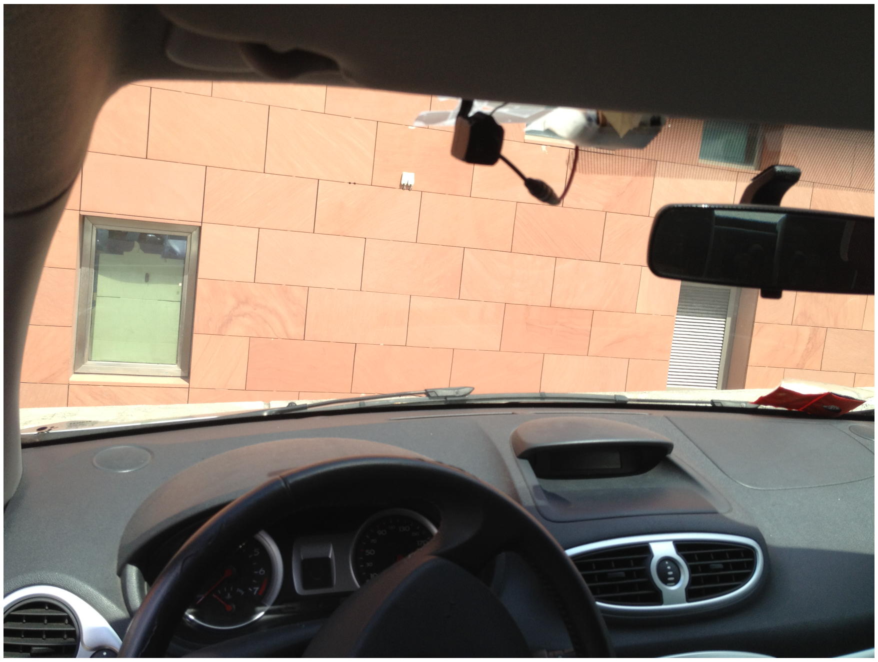

Figure 2: **The wi-fi camera set.**

As detailed in the previous section, the proposed system is composed of three processing blocks , fig. 1.

In the first block each video frame (a 640x480 pixels gray image, see figure 3) is passed to the Viola-Jones detector. The resulting region of interest is cropped around the center of the region, giving rise to a small frame of fixed dimension (280x280 pixels). If the detected ROI is smaller then 280x280 pixels, remaining pixels are set to zero. Some samples of extracted ROIs are shown in figure 4.

In the second processing block, the Sanger neural network takes in input the extracted ROIs and computes a small feature vector.

Starting from a typical number of principal components (12) used in eigen-faces detection [17] and using a small number of training frames (36 frames, representative of different positions of the head and different levels of attention) we found the best configuration for 16 principal component. Once fixed the weights of the hidden Sanger layer, data reduction can be easily obtained by projecting each ROI in the final feature space (i.e. by product of the Sanger weight vector for the row data frames). It is worth noting that this operation is very fast, giving as a result a very compact representation of each ROI.

For the third processing step, the definition of the number of neurons in the hidden layer has been obtained by repeating the training and test procedure for various configurations. We found optimal results for a hidden layer composed of 9 neurons.

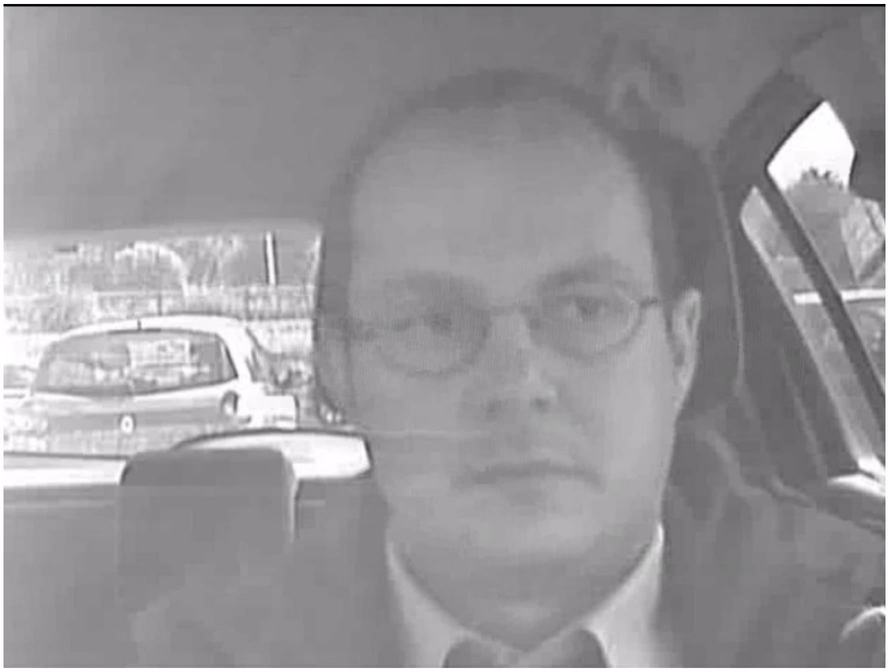

Figure 3: Example of a single frame extracted from the b/w mpeg2 video acquired by the wi-fi camera.

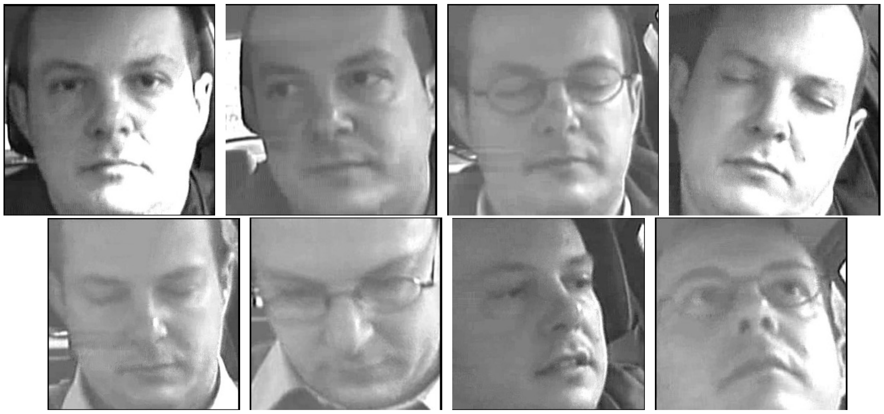

Figure 4: Some samples of 280x280 pixels ROIs extracted by the Viola-Jones detector. From the top left: two normal attention states, two inattentive states due to fatigue and three inattentive states due to distraction.

Table 1 . Classes distribution in the sets

Class	Train	Validation	Test
0-Attentive	817	821	8580
1-Inattentive- fatigued	917	909	9752
2-Inattentive- distracted	211	215	2136
Total	1945	1945	20468

Table 2. Results and accuracy mean weighted on the sets after the feed forward back propagation network FF-Bp

Class	**Validation set**	**Test set**
0-Attentive	95.52%	93.28%
1-Inattentive- fatigued	89.11%	88.02%
2-Inattentive- distracted	71.63%	74.72%
Acc. Mean weigh.	89.46%	88.84%

4.2 Database

The experimental database has been prepared starting from the available 6 sessions, for a total of 18 minutes of video and 29211 ROI frames. In order to guarantee performance and data reduction, all the ROIs have been coded in the Sanger feature space, as described above.

It is well known in the pattern recognition community that a crucial step in the experimental phase concerns the identification of three different sets of data (training set, validation set and test set). In fact, a good random distribution of the samples in these data sets guarantees a correct measure of the system performance, compensating for possible biases. To this aim, we used all the dataset of 29211 Sanger vectors to train a Self Organizing Map [18] having 16 input and 4x4 neurons in the Kohonen layer. This unsupervised neural network provide an equilibrate mapping of the dataset for the random sampling.

Table 1 shows the overall distribution of the resulting data sets. Note that the number of samples of the class 2 is not comparable to that of class 0 and 1 due to well known limitations of the Viola Jones algorithm (if the rotation of the head is significant the VJ algorithm does not return a ROI) [7].

4.3 Results

The training and the validation sets are used for the configuration phase of the feed forward network. Table 2 outlines the results obtained applying this classifier to the validation and test sets.

Table 3. Results in terms of normal situation or alarm on the sets after the feed forward back propagation network FF-Bp

FF-Bp 16-9-3	**Validation set**	**Test set**
Normal	95.52%	93.28%
Alarm	91.19%	90.97%
Acc. Mean weighted	92.60%	91.94%

If we express these results in terms of normal driving situations (attentive driver) or alarm situations (inattentive driver) the mean weighted accuracy raises up to 92% (table 3).

5 DISCUSSION AND CONCLUSION

The automatic system presented in this paper performs very well in the detection of the attention state of a single driver, in different environmental conditions. The outlined procedure does not use deterministic algorithms, nor specific feature-based approaches. Instead, a trainable iconic approach, based on neural network classifiers, is proposed. This method has several interesting characteristics:

it allows for a simple generalization of the attention states: additional states can be easily introduced and learned from a relatively small training set;

it allows for simple and straightforward improvement of the performance by adding new training samples.

It guarantees fast response; after the first training of the system, the global processing time is essentially the the time required for the Viola-Jones ROIs detector; simple vector algebra (Sanger coding and FF-Bp classification) cannot add significant extra time.

Concerning weak points, it is worth noting that an initial training of the system is so far required for each new user; this procedure requires about 3 minutes of training, which is an acceptable duration, but also requires an active cooperation of the new user, who must simulate both attentive and inattentive states. Current research is devoted to the simplification of this first training phase, making use of general models, totally independent from the single user, and a minimal training procedure of about 5 seconds . The approach is also suitable for the introduction in the vehicle of a face recognition based security system. In fact the same minimal training procedure can be easily used to code and then recognize the driver without errors from a small set of enabled users.

6 REFERENCES

1. A. Kircher, M. Uddman, J. Sandin, “Vehicle control and drowsiness” Technical Report VTI-922A, Swedish National Road and Transport Research Institute ,

2002.

2. The role of driver fatigue in commercial road transport crashes. European Transport Safety Council, Brussels 2001.

3. L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and E. Lopez. Visual Monitoring of Driver Inattention. In Comput. Intel. In Automotive Applications, SCI, pp 25-51, Springer, 2008.

4. W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, Face Recognition: A Literature Survey, *ACM Computing Surveys*, pp. 399-458, 2003.

5. J.P. Batista. A Real-Time Driver Visual Attention Monitoring System. In Lecture Notes in Computer Science: Pattern Recognition and Image Analysis, vol 3522, pp. 200-208, Springer, 2005.

6. S. Singh, N.P. Papanikolopoulos. Monitoring driver fatigue using facial analysis techniques. In Proc. Of the IEEE Int. Conf. On Intelligence Transportation Systems, pp. 314-318, 1999.

7. R. Senaratne, B. Jap, S. Lal, A. Hsu, S. Halgamuge, and P. Fischer. Comparing two video-based techniques for driver fatigue detection: classification versus optical flow approach. Mach. Vision Appl. 22, 4. July 2011.

8. black R. Senaratne, D. Hardy, B. Vanderaa, and S. Halgamuge. Driver Fatigue Detection by Fusing Multiple Cues. In Proceedings of the 4th international symposium on Neural Networks: Part II--Advances in Neural Networks, Springer-Verlag, Berlin, Heidelberg. 2007.

9. H. Rowley, S. Baluja, and T. Kanade: Neural Network-Based Face Detection. IEEE Tr. On PAMI, January 1998.

10. K. Sung and T. Poggio. Example-based learning for view-based face detection. IEEE Tr. On PAMI, January 1998.

11. P. Viola and M. Jones , Robust Real Time Object Detection, *International Journal of Computer Vision*, Volume 57, Issue 2 Pages: 137 – 154, 2004

12. J. Kennedy, R. Eberhart. Particle Swarm Optimization. Proceedings of IEEE International Conference on Neural Networks. IV. pp. 1942–1948.(1995).

13. A.A. Lenskiy and J. Lee. Drivers Eye Blinking Detection Using Novel Color and Texture Segmentation Algorithms. Int. J. of Control, Automation and Systems 10(2). 2012.

14. Paul Smith, Student Member, IEEE, Mubarak Shah, Fellow, IEEE, and Niels da Vitoria Lobo. Determining driver visual attention with one camera. IEEE Tr. On Intelligent Transportation Systems, Vol. 4, n. 4, Dec 2003.

15. O. Duda, P. E. Hart, D. G. Stark, “Pattern Classification“, second edition, *A Wiley-Interscience Publication* John Wiley & Sons, 2001.

16. T. D. Sanger , “Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network”, *Neural Networks*, vol. 2, pp. 459-473, 1989.

17. M. Turk, A. Pentland, “Eigenfaces for recognition”, *Journal of Cognitive Neuroscience*, 71-86, March 1991.

18. T. Kohonen, Self-Organizing Maps, Springer, 2001.